Technological Complements to AI Growth

Daniel Rock
Wharton, U. Penn

Prasanna (Sonny) Tambe
Wharton, U. Penn

Zhiwei (Berry) Wang*
Wharton, U. Penn

June 16, 2022

DRAFT. Please do not post, quote, or circulate.

Abstract

We provide theory and evidence connecting successive waves of IT investment to the marketplace advantages enjoyed by large, digitally intensive firms. We present a model requiring firms to make sequential tech investments to move up the AI "stack" and to invest in new organizational capital at each stage. We extend growth accounting techniques to estimate productivity growth in producing intermediate technological goods within firms. Laggard firms incur additive adjustment costs when catching-up with frontier firms. Since firms use other technologies to produce frontier technological intermediate goods, advantages in earlier vintages of technology capital "stack" in producing frontier intermediates. We take this model to a new panel data set on firms' technology investments. We demonstrate a progression in stages of major technology investments over the last two decades – databases, networks, cloud computing, data science, and AI – and demonstrate that adjustment costs across stages contribute to a "Matthew effect" in which firms with more developed digital capabilities face lower costs of remaining at the technology frontier. Over time, advantages in older technologies dissipate.

Keywords: *AI, digitization, productivity, superstar firms, information technology, complements, market power, machine learning*

*Authors are listed alphabetically. Comments are welcome at rockdi@wharton.upenn.edu & tambe@wharton.upenn.edu & zwangcj@wharton.upenn.edu. We are grateful for valuable feedback from Ananya Sen and from seminar participants at the Workshop for Information Systems & Economics.

1 Introduction

Information technologies play a key role in explaining performance differences among firms, and they may be particularly important for explaining the growing economic importance of large platform-based firms such as Amazon, Apple, Microsoft, and Google. These technology platforms have become particularly powerful over the last decade, in part, by leveraging data, algorithms, and AI to drive new business innovation.

Yet, why these technologies drive competitive advantage remains puzzling because much of the software needed to implement machine learning algorithms is freely available and computing resources are readily available through the "cloud". This raises the question: why have larger, tech-forward firms derived a disproportionate amount of the value from investments in data, algorithms, and machine learning? Although tech firms are now playing an outsized role in the economy, there is still much we do not know about the forces that are driving this concentration of economic activity. Most work in this area focuses on demand side forces (Eisenmann et al., 2006), but there are open questions about the relationship between technology and firm scale.

One potential explanation for this phenomenon can be found in the literature on business transformation and IT complements. Re-engineering business processes to support new technological investments can be difficult and expensive. Prior work has shown that heterogeneity in business processes explains a great deal of variation in the success that firms experience when adapting to new information technologies.[1] However, why these investments are easier for some firms to make remains an open question; it implies that a subset of firms enjoys advantages in the markets for resources needed to build the workplace complements required to support investments in AI and related technologies.

In this paper, we present theory and empirical evidence connecting firms' movements up the AI "stack" – the series of technology investments leading to AI-enabled prediction and comprised of prior investments in networks, databases, big data, cloud computing, data science, and finally, AI – to a "Matthew effect" wherein firms further along the stack face lower adjustment costs when moving to the next level of technological capabilities.

We first present a model where technological complementarities exist *within* vertical categories of input technologies as well as *between* technologies and business practices (e.g. IT and sales, generated by adjustment of work practices for each level of technology in the stack). Returns to investment in a technology are determined by the extent of investment in related technology input categories. Within this framework, it becomes increasingly difficult for technology laggards to catch up with firms at the technology frontier.

In the main part of our analysis, we test this theory using a rich, longitudinal panel data set of firms' technology investments. The data we analyze capture firm's investments over decades, from personal computing to networks and databases through more recent investments in data science and machine learning. A key strength of this analysis is the ability to empirically differentiate the

[1] For example, see Bresnahan et al. (2002); Susarla et al. (2011); Bardhan et al. (2013).

timing and intensity of firms' investments in these different IT vintages over the last twenty years. In contrast, note that most consistent, firm-level panels of technology investment have focused on aggregate IT stock or individual technologies (e.g. EMR or SAP systems). Panels that capture how firms react to different waves of technology investment have been elusive. This may be why there is no other work, to the best of our knowledge, that tracks the implications of firms' technological trajectories. Technology investments in our analysis are measured by examining firms' employment of technology workers with different skill endowments, which is similar to an approach that has been used prior work (Tambe and Hitt, 2012). The strengths and limitations of this approach are discussed later in the manuscript.

Our first set of descriptive results show that in our large sample, firms tend to invest in elements of the AI technology stack – networks, data, the cloud, data science, and AI – sequentially, consistent with lexicographic dependence. In other words, investments in earlier (older) technologies are important for successful investment in emerging technologies. This is consistent with the presence of additive adjustment costs for technologies that occupy a positions further up the stack, and suggests the difficulties firms face when attempting to "leapfrog" to new technologies. For instance, firms that have not made the appropriate investments in data collection and engineering cannot begin to build AI capabilities that require significant data assets. These results echo survey data results on business intelligence capabilities reported in Zolas et al. (2021), who demonstrate the progression of a large sample of firms through different stages of this technology stack.

We then turn towards examining the relationship between investments in these technology vintages and firm performance, using productivity regressions. Productivity regressions express engineering relationships between firms' factor investments and the output that a firm generates, and these econometric approaches have been central to much of the literature on the relationship between IT and productivity. To test these relationships, we create a panel data set on firms' investments in production factors (labor, non-IT capital) and investment levels in major technology categories. Using this econometric approach, we demonstrate that the success of each stage of the IT investment stack relies on complementary investments in earlier technology vintages. Correlations between productivity and AI are completely explained by firms' investments in data and data science, for example. This empirical approach reveals a firm-level "technology skills tree" that delineates the pathways firms take in unlocking new and advanced algorithmic capabilities.

Furthermore, we show that this dynamic investment pathway drives economic divisions among firms. These relationships allow firms with richer technological foundations to grow their technological capabilities more quickly than other firms, such that firms in the top quintile of technological assets are much more successful with AI than firms in the other quintiles.

Although several mechanisms may be at play, we focus in particular on one - skilled technological labor. We show that firms with more up-to-date technological investment are more successful at attracting workers skilled in more recent technologies. With a relatively tight market for skilled workers, this type of sorting makes it difficult for laggard firms to attract the workers needed to update their technological capabilities. Moreover, the rapid turnover of these technologies means

that the labor market is rapidly updating their skills (Chari and Hopenhayn, 1991), and moves beyond firms that lag too far behind the frontier.

Our key finding, then, is that firms' digital capital begets additional digital capital. This finding is connected to two separate literatures. First, we contribute to the existing literature on IT and performance by considering how successive waves of IT investment endogenously generate divisions in the performance of firms. Although the empirical literature on IT investment performance is large and mature, we are not aware of work that differentiates how successive waves of IT investment affect performance.

Second, we connect the trajectory of these investments to evidence of increasing returns in the digital economy. We also, therefore, contribute to a growing literature documenting greater levels of market power in a small number of firms. These papers have documented the emergence of so-called "superstar" firms and some of their consequences (e.g. price markups) (Autor et al., 2020; Hall, 2018; Van Reenen, 2018; De Loecker et al., 2020), and recent work provides evidence that some of these differences are rooted in digitization (Covarrubias et al., 2019; Tambe et al., 2020). However, we are not aware of other work that uncovers the joint technological and organizational mechanisms that contribute to these concentration effects, or that explore supply-side forces between technology and market concentration.

2 Related literature

2.1 IT investments and performance variation

A defining question in the IS literature, since the 1980's, has been uncovering the role of IT investment in driving performance gains for firms. Although evidence of this connection was scarce in the 1980's and early 1990's, by the late 1990's and early 2000's scholars had begun to provide a substantial body of evidence, using a number of data sets and approaches, demonstrating robust correlations between investments in information technology and different measures of corporate value, including productivity (Barua et al., 1995; Brynjolfsson and Hitt, 1996; Dewan and Min, 1997), market value (Bharadwaj et al., 1999), and profit (Mithas et al., 2012). Dedrick et al. (2003) provides a comprehensive review of this literature.

One limitation of this literature is that it focuses on a specific class of technological investments made in the late 1990's, largely those related to personal computing and networked computers. These investments in networked communications and personal computing do not capture the subsequent waves of technological investments in Internet communications, data collection and data science, or most recently AI. A few papers extended analysis of the economic effects of aggregated technology investment into more recent time periods (Tambe and Hitt, 2012), but with only a few exceptions (Bloom et al., 2014), have not examined the effects of different technology investment categories. This literature, therefore, does not address the fact that different categories of IT investment can have differing implications for the most effective ways to organize work.[2]

[2]Other literature in IS examine the impact of specific technologies like SAP and ERP systems on performance. This

A central question pursued in this broader literature relates to the variation in performance gains that firms realize from their IT investments. An early finding from the empirical literature on IT and productivity was that the dispersion of returns from IT investment was very large (Brynjolfsson and Hitt, 2000). One of the most important explanations for this variation is that firms require significant "co-invention" to make IT investments productive (Bresnahan et al., 1996). The development of these intangible complements through this co-invention process can take time, leading to empirical regularities in the measurement of productivity derived from these investments (Brynjolfsson et al., 2021). Firms differ in the way in which they organize their central work processes, and this variation is important for explaining different benefits from IT investments. When firms do not have appropriate work practices in place, they have to undergo an expensive re-engineering process (Hammer and Champy, 2009), with costs on the order of ten times that of the technology investments themselves (Saunders and Brynjolfsson, 2016).

The theoretical foundations for this argument can be found in the literature on supermodularity in business practices (Milgrom and Roberts, 1990; Brynjolfsson and Milgrom, 2013), which argues that firms are better off when information technologies are installed along specific business practices. This line of reasoning argues that these investments are more than the "sum of their parts". This argument has a robust empirical foundation. Existing econometric studies integrate data on information technology investments and work practices to demonstrate that firms realize higher IT investment returns when the work practices they adopt are complementary to these technologies (Black and Lynch, 2001; Brynjolfsson et al., 2002; Caroli and Van Reenen, 2001; Bartel et al., 2007; Tambe et al., 2012). Most often, these studies have focused on work practices identified in the literature on high-performance work (Osterman, 1994; MacDuffie, 1995; Batt, 1999), which include, for instance, front-line discretion, teamwork, and problem-solving.

This literature leaves open a number of important questions. For our purposes, an important omission in this literature is that it analyzes the relationship between IT and productivity, but not distributional effects or why some firms derive long-term advantage from technological progress (see Dedrick et al. (2003) for a similar argument). The current literature in this area is silent on how firms may differ in their trajectories of investment (or their investment returns).

2.2 Climbing the AI ladder

This omission seems particularly important now as firms that have historically been at the frontier of digital investment appear to be moving along a different trajectory than other firms that have lagged behind.

The last decade has borne witness to a surge of corporate interest in technologies that convert data to superior decision-making (Manyika et al., 2011). Since the wave of investment in e-commerce that occurred during the dot-com boom, firms have been collecting growing amounts of data on

literature also dives into the mechanisms through which these different technologies are likely to drive performance, but are largely separate from the theoretical frame we pursue here, which focuses on the general-purpose nature of IT investments that can transform decision-making within the organization.

everything from consumer preferences to financial transactions. A focus of the last decade has been on implementing technologies that can aid in converting these massive databases into effective and profitable predictions. This includes investments in technologies that are useful for storing and sharing data, such as cloud-based technologies (Jin and McElheran, 2017), and for processing large data sets, such as decentralized data technologies such as Hadoop and Spark (Tambe, 2014). The most recent work in this area focuses on data science, machine learning, and AI technologies (Rock, 2019; Acemoglu et al., 2020).

Much of the reason for the continued interest in this question is because information technology, like electricity, can be viewed as a general-purpose technology, capable of driving productivity gains across all economic sectors, and AI investments have the potential to continue to drive this type of economic transformation (Goldfarb et al., 2019). Firm-level analyses have demonstrated that investment in data science and data analytics can drive productivity and innovation (Wu et al., 2020). Moreover, early evidence suggests that investments into AI can change how decisions are produced and have documented value returns from these changes (Abis and Veldkamp, 2020; Rock, 2019).

Key to our argument is that investments in AI require capabilities at several layers of the AI technology "stack" (Zolas et al., 2021). The key input into AI decision-making is data. Effective AI investment requires having incurred adjustment costs at all levels of this stack, beginning with extensive data collection, storing, sharing, and sharing, progressing to large-scale processing and data science, and then finally, to machine learning technologies which can automatically make decisions when provided with data in the correct format. Firms that have not made the appropriate investments at lower levels of this stack cannot make effective higher order investments. For instance, brick-and-mortar stores that never expanded into e-commerce may not have collected the volumes of data needed to train machine learning models. Before implementing AI capabilities, these organizations will have to incur the adjustment costs for earlier stages such as data collection and digitization, which can be extensive.

In this paper, we argue that investments at different levels of this stack exhibit sequential dependence. That is, investments in technologies further up the stack only yield productivity benefits if technologies at lower levels of the stack have been installed in the firm. Although the implied dependencies are likely to be true of the technologies underlying many types of capabilities, the transformative and general-purpose nature of AI makes this relationship particularly important, and connects it to broader economic divisions. Firms seeking to climb the AI ladder have to navigate both the co-invention and changes in business practices required at each layer of the stack as well as the "vertical" complementarities between different technology layers as firms move further up the stack.

2.3 AI, intangibles, and superstar firms

A final literature from which this paper draws concerns the growing levels of economic concentration in superstar firms. Recent work has focused on the rising levels of market concentration

occurring in many industries (Van Reenen, 2018; Autor et al., 2020). These papers illustrate that a small number of mega-firms are responsible for an ever-larger amount of economic activity. The associated change in market power has a variety of implications, ranging from monopsony power in hiring to increased pricing power in product markets (Hall, 2018). Some have connected this change to the disproporationate accumulation of intangible capital in some firms, including those intangible assets associated with growing levels of technological investment (Bessen, 2020; Crouzet and Eberly, 2018; Covarrubias et al., 2019). Here, the relationship is that economies of scale in intangible capital drive greater value in firms in which these assets have accumulated. These assets allow firms to rapidly scale innovation and logistics across markets.

The adjustment costs necessary to implement AI technologies – both those associated with the different technologies needed as well as the co-invention needed for each stage of investment – present a barrier to digitizing firms, especially those that have not made the investments that form the foundation of this technology stack. The adjustment costs associated with a full installation of "intangibles" are likely significant and may have implications for competition and for industry structure.

This raises the question of what the competitive barriers may be that allow some firms to scale in this way. This paper argues that the requirements to climb the technology stack in the way described above, along with investment in the concomitant work practices, can endogenously generate these "rich-get-richer" effects wherein firms with developed digital capabilities face lower adjustment costs when moving onto the next phase. In the next section, we formalize these ideas.

3 Technological complements

3.1 Theory

We refer to complements across functions as "horizontal" complements, while complements within a given organizational function are "vertical" complements. Estimation of performance complementarities in information systems often rely on either performance regression tests or measuring correlation of business investments. Applications of these tests in the information systems literature focus on measurement of aggregate complementarities.

To uncover the mechanisms driving recent market concentration, it may nevertheless be of interest whether returns to complementary investments arise from horizontal interactions across business functions or from vertical supermodular performance gains in deepening a particular functional capability. In particular, we might expect similar returns to a variety of possible IT investments from horizontal complements (e.g. legal departments are unlikely to have strong opinions about the use of python or R). On the other hand, specific technologies with vertical complementarities may rely on simultaneous investment in other specific technologies. Differentiating between these two sources of supermodular performance gains is an extension of the complements literature and is of interest to managers evaluating and implementing new technology investments. As a simple example, consider a Cobb-Douglas production function $Y = AF(K,L)$ with output Y,

total factor productivity A, labor input L, and aggregate capital input K such that:

$$Y = AL^{\alpha}K^{(1-\alpha)} - wL - rK \tag{1}$$

α denotes the factor share of labor and shares sum to 1. w and r denote the wage and capital rental rate (respectively). Now, let K be determined by a nested constant elasticity of substitution (CES) production function of two varieties of capital K_1 and K_2:

$$K = (\theta_1 K_1^{\frac{\sigma-1}{\sigma}} + \theta_2 K_2^{\frac{\sigma-1}{\sigma}})^{\frac{\sigma}{\sigma-1}} \tag{2}$$

In this capital aggregator, σ denotes the elasticity of substitution parameter and θ_i is the productivity of each capital variety for $i \in \{1, 2\}$. We then have the following equation for the marginal horizontal complementarity in 3 below if the marginal product of either factor equals its price (w or r):

$$\frac{\partial^2 Y}{\partial K \partial L} = \frac{(1-\alpha)w}{K} = \frac{\alpha r}{L} \tag{3}$$

and then the marginal vertical complementarity in producing aggregate capital K is:

$$\frac{\partial^2 K}{\partial K_1 \partial K_2} = \frac{\theta_1 \theta_2}{\sigma} \left(\frac{K^{2-\sigma}}{K_1 K_2}\right)^{\frac{1}{\sigma}} \tag{4}$$

This special case illustrates that complementarities across the aggregate capital and labor inputs are declining in each respective input (that is, the cross derivative of the production function is positive and smaller at larger levels of capital or labor). This function would apply equally to changes in K_1 or K_2. At the same time, the vertical complementarities in 4 are weakly positive, but depend on how substitutable each component input is in creating the aggregate capital quantity. We can estimate the source of supermodular performance as horizontal or vertical by generating a hierarchy to sort inputs by category.

3.2 Growth Accounting with Contingent Vertical Complements

In the case of AI, the presence of a particular input to the production function relies upon the use of other investment inputs such as computing, data asset investment, and data science. These inputs are a special kind of vertical complement such that production of the AI asset for use in the overall production function is a contingent necessity. That is, without data science investment, there can be no AI investment (but not necessarily vice versa). More generally, assume there is some capital input quantity $K_A I$ for firms that must be produced using other capital input quantities K_i and labor units L where i indexes capital types. In ordinary growth accounting we have:

$$Y = AF(\mathbf{K}, L, K_{AI}) \tag{5}$$

where Y is output priced at the numeraire, F is the production function, A is total factor

productivity, and $\mathbf{K}$ is a vector of capital quantities.

$$\frac{dY}{Y} = \frac{dA}{A} + \frac{wL}{Y}\frac{dL}{L} + \sum_{i=1}^{I} \frac{r_i K_i}{Y}\frac{dK_i}{K_i} \tag{6}$$

This is the standard growth accounting equation with factor prices w for labor and r_i for each variety of capital i. With $g_x = dx/x$, we solve for the residual as:

$$g_A = g_Y - \frac{wL}{L}g_L - \sum_{i=1}^{I} \frac{r_i K_i}{Y} g_{K_i} = g_Y - \frac{wL}{L}g_L - \boldsymbol{\alpha}'\mathbf{g_K} \tag{7}$$

In this case $\boldsymbol{\alpha}$ denotes the column vector of capital factor shares and $\mathbf{g_K}$ is the vector of capital growth rates. Now let $\omega G(\mathbf{K}, L)$ be the internal production function for AI output using capital and labor inputs of other varieties, priced at r_{AI}. ω is the firm's total factor productivity *in producing AI*. Denoting any input quantity X used toward producing AI as X^{AI} to distinguish it from the total quantity of that input X used by the firm overall, we then have:

$$r_{AI}K_{AI} = r_{AI}\omega G(\mathbf{K}, L) \tag{8}$$

$$\frac{dK_{AI}}{K_{AI}} = \frac{d\omega}{\omega} + \frac{wL^{AI}}{r_{AI}K_{AI}}\frac{dL^{AI}}{L^{AI}} + \sum_{i=0}^{I} \frac{r_i K_i^{AI}}{r_{AI}K_{AI}}\frac{dK_i^{AI}}{K_i^{AI}} \tag{9}$$

or more simply,

$$g_{K_{AI}} = g_\omega + \frac{wL^{AI}}{r_{AI}K_{AI}} g_{L^{AI}} + \boldsymbol{\gamma}'\mathbf{g_{K^{AI}}} \tag{10}$$

As in the case of $\boldsymbol{\alpha}$ above, $\boldsymbol{\gamma}$ denotes the vector of factor shares of each capital variety in producing AI output. This equation then decomposes growth in AI capital as a function of growth in AI capital production productivity and growth in inputs required to create AI. Note that this is in no way specific to AI capital, but rather fits the description of any downstream internally produced capital input.

Equation 10 can be estimated to recover AI capital productivity growth with a cross-section of firm-level data. There are a few practical requirements in estimating the rate of AI productivity growth g_ω. The first is that AI capital prices and quantities must be measurable, and the second is that the proportion of total inputs devoted to AI production must also be measurable. In practice this may prove difficult. To overcome these hurdles we propose two approaches. The first is to assume X^{AI} for all input types is a fixed constant proportion of overall inputs. $X_i^{AI} = \lambda X_i \; \forall \; i$, with $\lambda \in [0,1]$. That leads to a similar equation to the above in equation 10, but a different interpretation of the coefficients of a regression of capital quantity growth in AI on the growth of other inputs:

$$g_{K_{AI}} = g_\omega + \lambda\frac{wL}{r_{AI}K_{AI}} g_L + \lambda\boldsymbol{\gamma}'\mathbf{g_K} \tag{11}$$

Estimating a panel regression with firm fixed effects following equation 11 will recover the firm

average AI production productivity growth as the intercept, while coefficients on the growth rates of other inputs will reflect the AI product factor share of other inputs multiplied by the constant λ share of those inputs in producing AI.[3]

More natural than assuming a constant share, however, might be to assume that the proportion of inputs devoted to producing a downstream contingent input type (i.e. a vertical complement) is proportional to the share of firm value accounted for by the input variety being assembled by other inputs. For the case of AI, define λ_{jt} as the time t proportion of firm j value that AI capital represents (P_{it} is the price of capital i at time t).

$$\lambda_{jt} = \frac{\sum_{t=0}^{\infty} r_{AI,t} K_{AI_{jt}}}{\sum_{i}^{I} P_{it} K_{it}} \tag{12}$$

Define $\mu_{jt} = \frac{\sum_{t=0}^{\infty} r_{AI,t} K_{AI_{jt}}}{r_{AI_t} K_{AI_{jt}}}$, the ratio of the total value of capital service flows from the downstream contingent vertical complement (AI in this case) at time t for firm j to the capital service flow from AI in period t. $V_{jt} = \sum_{i}^{I} P_{it} K_{it}$ is the value of the firm overall. Now we have:

$$g_{K_{AI_{it}}} = g_{\omega_{jt}} + \mu_{jt}[\frac{w_t L_{jt}^{AI}}{V_{jt}} g_{L_{jt}^{AI}} + \sum_{i=0}^{I} \frac{r_{it} K_{ijt}^{AI}}{V_{jt}} g_{K_{ijt}^{AI}}] \tag{13}$$

Valuations for AI in a firm can be recovered with hedonic regression techniques, yet estimating equation 13 also requires factoring growth in the AI share of value in the firm overall out from the growth in the units of capital input used in production for all purposes. That is,

$$g_{K_{ijt}}^{AI} \approx \ln(\frac{\lambda_{jt} K_{ijt}}{\lambda_{j,t-1} K_{ij,t-1}}) = ln(\frac{\lambda_{jt}}{\lambda_{j,t-1}}) + ln(\frac{K_{ijt}}{K_{ij,t-1}}) = g_{\lambda_{jt}} + g_{K_{ijt}} \tag{14}$$

Equation 14 defines growth in the components used to produce the AI investment variety as a function of growth in AI's value of the firm overall and growth in overall capital inputs used to create AI (and elsewhere in the production function). Subbing equation 14 into 13, we find:

$$g_{K_{AI_{jt}}} = g_{\omega_{jt}} + [\mu_{jt}][g_{\lambda_{jt}} + \frac{w_t L_{jt}}{V_{jt}} g_{L_{jt}} + \sum_{i=0}^{I} \frac{r_{it} K_{ijt}}{V_{jt}} g_{K_{ijt}}] \tag{15}$$

Each of these constituent growth rates are measurable; The interpretation of the coefficients on growth rates in a regression for equation 15 will therefore be the factor share contributions of inputs used in constructing AI for later use in production multiplied by a common coefficient representing the ratio of AI market value to current capital service flows. The growth rate of AI capital market value should therefore be included as a control to back out factor shares. The firm-specific intercept of a model with firm fixed effects will reflect the firm-level time average of AI production productivity growth, while a model with time fixed effects recovers the cross-sectional average of AI production productivity growth instead if the model is properly specified (especially

[3] λ can in principle vary by capital or labor variety as well, in which case the $\boldsymbol{\lambda}$ vector of input weights must be multiplied element-wise by the elements of $\mathbf{g_K}$.

with regard to omitted variables, which will incorrectly be absorbed by this total factor productivity in producing AI measure if not properly included). The relationship between AI-producing total factor productivity growth and overall total factor productivity growth will straightforwardly be the factor share of AI capital service flows multiplied by the growth of AI-producing total factor productivity:

$$\frac{r_{AI,t}K_{AI,t}}{Y_t}g_{\omega_{jt}} \tag{16}$$

From this expression it is clear that total factor productivity in producing AI may grow substantially if $g_{\omega_{jt}}$ increases. Yet overall total factor productivity might not be affected by AI at all if the contemporaneous capital service flows from AI are still zero. Nevertheless, productivity growth in producing AI is still measurable if AI quantities and other inputs are observed. We will estimate the relative contributions of related skills in the growth of AI capital within firms and recover an estimate of total factor productivity growth in producing AI.

3.3 Hypothesis development

We consider the case that firms have to do internal production with and/or reconfiguration of marketable inputs to create the ingredients necessary to produce final goods and services. This intermediate production step is often where new technologies operate. If a new technological input variety is hard to purchase externally but still valuable in final goods production, then the firm must build that intermediate new technology internally using contingent *vertical* complements as ingredients. And even in the case that the intermediate technology is marketable, it might still be advantageous for firms to make rather than buy it. The foregoing theoretical sections have a number of implications for the empirical regularities that might accompany such a production structure, particularly for the case of AI and ML. AI technologies remain difficult to purchase on the open market in shovel-ready formats. Even in the case that off-the-shelf algorithms are deployed, our data suggest that many companies still choose to hire data scientists and machine learning engineers. Even if those ML-employing firms use standard software, these hires are still valuable if only to make the necessary adjustments for productive deployment of software. Prior waves of information technologies have demonstrated similar trends in the evolution of adjustment costs (Tambe et al., 2020).

If a new technological variety like AI is valuable in production, then the first empirical evidence we should find is that part of the firm's market value should be strongly correlated with the presence of that technology. For the example of AI, not only might AI capital be valuable, but like other IT assets the complementary business processes and intangibles associated with AI production will also correlate strongly with the market value of the firm (controlling for obvious other omitted variables in producing AI like data science, cloud computing, and other digital capabilities). We therefore directly test the hypothesis that AI assets have a positive price with hedonic regression of the form below:

$$V_{jt} = \sum_{i=0}^{I} \pi_i K_{it} + \delta' \mathbf{X_{jt}} + \epsilon_{jt} \tag{17}$$

where V_{jt} denotes the market value of firm j in year t, i indexes capital types as represented by skills counts (e.g. AI, Data Science, Cloud Computing, etc.), and X_{jt} represents a vector of controls. What we are explicitly testing is the value of π_i, the price on the different kinds of capital varieties K_i. Our hypothesis is that π will be positive and statistically significant for AI assets even conditional on the presence of other kinds of related assets. We also expect that these coefficients on value might not be stable – entry and exit dynamics in the marketplace and accumulation of related assets by different types of firms over the life cycle of a given technology might cause the market price represented by the hedonic regression coefficients to change.

π will include both the replacement cost of the assets as well as the adjustment costs to make new assets valuable to production processes (i.e. quasi-rents accruing to firms that use resources more efficiently than competitors). This replacement cost has an obvious interpretation for marketable technologies as the simple cost of purchasing new capital. In the case of an internally produced asset, it represents the input costs that are paid for externally procured or rental capital and labor. While our first hypothesis is therefore that the assets are priced as soon as firms devote significant effort to producing new technologies as a variety, as the adjustment costs and adoption dynamics for incorporating a technology like AI into production change we must expect that the price associated with that technology will change as well. We investigate these results by running cross-sectional regressions within year on both balanced and unbalanced panels to detect possible price changes Adjustment costs are therefore important at *each stage* of intermediate technology production. If data science assets and cloud assets must both be used to create AI within an organization, competitive advantages in deploying those assets will compound to make AI asset production cheaper as well. In a regression of market value on AI as well as its constituents, the AI component will absorb all AI-related adjustment cost quasi-rents. That includes competitive advantage in using AI, and also cost advantages in producing AI, cost advantages in deploying data science, and cost advantages in deployment cloud (for example). The residual coefficients on those constituents for AI relate to the Tobin's q value not associated with AI construction or deployment. Estimating the component of market value represented by AI assets is additionally useful for structurally recovering the total factor productivity in AI production present in equation 15.

A third series of hypotheses then concerns the productivity contributions of new technologies rather than the more direct pricing of AI assets. Productivity gains need not coincide with investment and may take significant lags to show up in measured statistics (Brynjolfsson et al., 2021). Nevertheless for mature technologies we should expect that production function estimation should recover some component of productivity growth that can be explained by technological investment (in our case proxied by labor flows). We hypothesize that sufficiently mature technologies will explain variation in output (with positive coefficients) and that coefficients on interactions between technology types will also be positively correlated with output.

At this early stage, it is not clear that AI should be a strong correlate of firm-level productivity. A number of papers have failed to find any meaningful correlation between contemporaneous measures of productivity and AI so far (Rock, 2019; Babina et al., 2020; Alekseeva et al., 2020; Acemoglu et al., 2020). Nevertheless, historically productive firms are more likely to have the necessary internal capital and labor assets required to produce AI. This filtration whereby the most productive firms invest in a given technology input vintage, then the more productive members of that cohort continue to invest in new vintages as they arise and then those vintages are used for subsequent technologies leads to a narrowing of the competitive field. The selection of firms into narrower categories of full-stack technology adoption might arise as a natural consequence of firms with higher overall levels of productivity facing stronger incentives to maintain a position at the technological frontier. This Matthew Effect (Merton, 1968) of AI – where large technologically sophisticated firms grow relatively more technology-intensive and sophisticated – is what we hypothesize to see in the early stages of AI investment. Because the toolkit, skills, and business processes to build AI capital are scarce and contingent upon the presence of antecedent technologies, AI will be more commonly found where other recently successful applications of information technologies are more prevalent. We test this hypothesis as well, examining whether or not larger and more productive firms in general also invest more heavily in AI and AI-related complements. This hypothesis test is operationalized as a productivity decomposition following equations 10 and 13 by regressing AI growth on growth of other measurable inputs. A positive and statistically significant correlation between AI growth and other related capital inputs (measured as skill count growth) provides suggestive evidence of a particular form of capital deepening: technological capital applied to create AI capital varieties.

Lastly we study the relationship between total factor productivity growth and total factor productivity growth *in producing AI*. We expect that the relationship should be roughly proportional to the factor share of AI in producing output.

4 Data

The data set used for this analysis is a panel of workforce capital stocks and stock market valuation data of 1,733 firms from the period of 2008 to 2020. We briefly describe the data sources below.

4.1 Technology investment data

Our technology investment data are constructed from workforce skills data. Our workforce data are from a partnership with Revelio Labs, a company focused on developing databases to support workforce intelligence. These databases are constructed from a variety of different data sources, including online profiles, job listings, and federal databases. The data we obtain from Revelio are similar in informational content to that posted on online professional networks such as LinkedIn. The data posted on online professional networks can be used to construct panel measures of human capital stocks and flows at the firm-level. For instance, one might measure a firm's investment

in engineers with specific skills, including those related to the cloud, data science, or machine learning.

In the academic literature, similar approaches have been used to measure technology investments by occupational counts (Lichtenberg, 1995; Tambe and Hitt, 2012; Tambe et al., 2020) as well as at the level of the individual skill (Tambe, 2014; Rock, 2019). We construct measures on firm's investment in different technology stacks using the sum of capital stocks of a basket of individual skills belonging to that stack. For example, AI is associated with "machine learning", "natural language processing", "image processing / computer vision", "artificial intelligence", "tensorflow", "pytorch" and "scikit-learn". The full list of skills and how they map to layers of the technology stack can be found in Appendix A.

Because IT employment is complementary to IT capital, we use these skill counts as a proxy measure for investment in technology. We relate skills to the firm's technology capital stock by making the assumption that a dollar of capital requires a fixed amount of labor with that skill to maintain it. That is, a firm that doubles its AI capital would require twice the AI labor to maintain it. This assumption allows us to use labor expenditure as a proxy measure for the firm's aggregate IT capital stock.

The principal challenge with using this approach is related to sampling. This is not an administratively generated data set, and so the numbers of workers with a skill are imputed by the data provider. This means there will be some measurement error in the firm-level measures of skills counts.

4.2 Market valuation and other data

The Compustat database was used to construct market valuation and related metrics used in the market value regressions. The created measures include total market value (market value of equity plus the book value of debt), Property, Plant, & Equipment (PP&E), Other assets (Total assets minus PP&E), R&D expenses and advertising expenses. This panel was merged with the technology data set. The merged data set has 19,101 observations over thirteen years for market value and technology stocks. There are 1733 unique firms represented in the data set.

5 Empirical Results

Figure 1 shows the imputed technology stack that AI firms require, derived from the Revelio skills data. In this illustration, each level of the technology stack is associated with a set of Revelio skills.[4]

[4]For example, AI is associated with the following skills: 'Artificial Intelligence', 'Machine Learning', 'Decision Trees', 'Natural Language Processing', 'Pattern Recognition', 'Speech Recognition', 'Random Forests', 'Support Vector Machines (SVM)', 'Neural Networks', 'Sentiment Analysis / Opinion Mining', 'Bayesian Networks', 'Bayesian Methods', 'Naïve Bayes', 'Bayesian Modeling', 'Markov Chains', 'Knowledge Discovery', 'OpenNLP', 'Principal Component Analysis', 'K-Means', 'Nearest Neighbor algorithm', 'Latent Dirichlet Allocation', 'Sentiment Classification', 'Kernel Methods'. For brevity, the skills associated with the other levels of the technology stack are not shown here, but are reported in Appendix A.

We apply an embeddings approach (detailed in the Appendix) to recover skills clusters associated with AI. The figure suggests a ladder where among firms that invest in IT, a subset of those invest in data related technologies. A subset of these invest in technologies related to networks. A subset of these invest in technologies related to the cloud. A subset of these invest in technologies related to AI. The grey lines that lead to AI indicate that a mixture of these investments lead to AI.

Additional descriptive results are presented in Table 1. Correlations among major technological categories indicate that these technologies co-occur inside firms. This is somewhat unsurprising, given the dependence of these technologies upon one another, but it is consistent with the argument that the different layers of this technology stack are complements in productivity.

The first step in our empirical analysis concerns the estimation of the market value associated with AI technologies. We find, as in Rock (2019), a positive and significant price on AI skills indices as well as other kinds of technology. The results for our primary logged market value decomposition as in equation 17 (but logging both sides) are reported in Tables 2 and 3 for balanced specifications with i) only year and ii) year and firm respectively. The coefficient on AI is positive and significant at a five percent level, controlling for other aspects of firm size. In particular, coefficients range between roughly 1.9 and 4.7 percent for the predicted change in market value given a 100 percent increase in AI talent. The full specification in column 9 assigns negative coefficients to some technologies like network and internet categories. These categories indicate older technology vintages wherein mentions of skill sets in those domains might indicate employment at a smaller firm, or alternatively, firms that are catching up to a vintage that has already matured. This provides evidence for our first hypothesis which is that construction of AI assets within the firm is valued by the market. Including firm fixed effects largely diminishes the positive correlation between AI and market value, suggesting that the relationship between firm value and AI is time-invariant and firm-specific.

The next stage concerns the change in this valuation of the AI and associated complementary assets over time. Tables 4 and 5 show the results of within-year regressions with similar specifications to the foregoing tables' column 9 specifications. AI skills are mostly not statistically significant in their valuations until recent years in the cross-section, though we note that these hedonic prices do not adjust for static firm-specific unobserved factors. As we observe a ramp-up in the value of AI from 2017-2020, we also see a growth in the coefficient on data science from roughly 2013 to 2020 (coefficients were positive and significant on data science prior to that, but the composition of data science skills also changed in the earlier years). Other technologies show little correlation in the stable balanced cross-section with market value, except for network skills. Network skills are negatively correlated with market value, suggesting that smaller companies are more likely to employ network specialists.

We first check that we can estimate productivity growth and check capital and labor coefficients against common benchmarks. Our results are in Table 6. The specifications show similar coefficients on capital and labor ("Emp" corresponds to employee counts). The constant time average productivity growth is roughly 2.3 percent per year across all firms, close to the estimates from

government statistical agencies.[5]

Our final set of regressions concerns the production of AI capital via the consumption of upstream vertical complements. This leads to the direct measurement of total factor productivity in building AI, one of the two terms that can increase output growth (the other is higher output share of AI capital services). We find evidence for the vertical complements story, with other skill varieties correlating strongly with AI growth.

We provide evidence for capital growth relationships within the technology stack by estimating a regression of firm's AI skills investment on complementary technology measures. The regression that we estimate is $\ln(g_{AI_{it}}) = g'_{X_{it}}\beta + \gamma_i + \nu_t + \epsilon_{it}$ where i indexes firm and t indexes the month (from January 2008 to April 2020). $g_{X_{it}}$ includes the indices for growth rates in software, cloud, and data science investments. We find evidence that the interactions of technologies like software engineering skills, cloud computing, and data science are predictive of AI skills intensity in specific skill combinations. Investment in AI is predated by investments in software and data science or software and the cloud.

Table 7 shows the AI production function on a balanced panel of firms (i.e. a panel in which firms cannot enter or exit the panel). This regression estimates the role of different technological factors in producing AI capital. Each successive regression adds a new layer of the AI stack. With all factors added, the correlations with big data and data science remain significant. This is a reasonable result because these two investments are most proximate to AI prediction. Notably, when relational database technologies are included directly in the model, the estimates on cloud measures are not significantly different than zero.

Table 8 estimates the same models using the larger unbalanced panel. Because it allows firm entry, highly digitally intensive firms can enter the panel. This table provides more robust correlations between investments in data, data science, big data, and other factors and the production of AI capabilities. Both of these regression tables stand in contrast to the overall output productivity results; AI capital growth is not correlated with overall output growth once we adjust for the growth of vertical complements to AI. This possibly indicates a pathway by which AI capital is built from the other technologies and then subsequently used to produce output alongside a low baseline capital service flow in the recent path from these AI capital investments. In the longer run, we expect the market value of AI capital to yield AI capital service flows, but in early stages of the technology it is possible to get market value without productivity enhancements.

In Figure 2 we show the HHI for different varieties of technology inputs over time. AI capital is much more concentrated than other kinds of technology. In Figure 3 we show the TFP in AI production quintiles after removing time means. That is, first we calculate the residual of a regression of AI growth on firm fixed effects and growth of other related technological skills types as above. Given de-meaned estimates of the productivity residual in producing AI, we have a noisy measure of the time-varying component of this AI TFP. We then group by quintile of AI level and

[5]These are slightly higher than the overall economy, but publicly traded firms tend to be more productive than the rest of the economy.

average to reduce the noise and get a coarse view of whether or not the time trend in AI TFP (in other words, how productive firms are in producing AI) reinforces growing differences between the most intensive AI capital builders and the least or if AI capital is becoming more equal. What we see is that the baseline growth rates over time in productivity for producing AI for AI-intensive firms are on average higher than low AI TFP firms, but that in recent years this gap has narrowed somewhat. It remains to be seen if the closing growth gap continues to narrow, but it is clear that for the recent past, inequality in AI capital accumulation has been growing. Generally it seems that AI-intensive firms do not have a special advantage in more productive use of AI inputs, but rather have a competitive advantage in AI capital input deepening that other firms do not have.

6 Conclusions and next steps

Our paper shows that the dependence of technology investments on the success of a firm's earlier investments has implications for rising levels of market power concentration. We provide evidence for this result by using novel and unusually granular data on firms' investments into different vintages of information technology. Among the top quintile of firms, this relationship across technology vintages contributes to faster levels of growth.

First, our findings have implications for scholars. Earlier work linking technology investment to productivity growth has principally focused on technology in aggregate stock measures or single waves of technology (e.g. networks, big data, cloud computing, EMR or SAP systems). This paper suggests the value of considering the path-dependence among these investments and underscores the importance of firm trajectory in explaining changing patterns of economic concentration. Although a significant body of work has measured the productivity of IT investments, more work is needed on how technological investment leads to long-run productivity differences.

Our study also has implications for managers. A major concern for modern managers is digital transformation. Managers in non-digitized industries often face significant competition from platform businesses that are moving into more and more business lines, and many that have sought to catch up with these platform businesses have struggled when it comes to building out digital capabilities. The path dependence highlighted in this study re-frames how managers must be thinking in order to build downstream AI capabilities. It suggests that managers of non-digitized firms looking to compete in an AI-enabled economy should be thinking about constructing the entire value chain, which includes all the concomitant business transformation that accompanies each of these steps.

There are, of course, a number of limitations to our study, and these provide some avenues for further work. Although the costs associated with each of these technologies is significant, economic forces are working to lower the costs of these stages, making it easier and cheaper for firms to adopt new technologies. For instance, no-code tools for data engineering make it relatively easier to build data capabilities for small to medium size firms that may not have the scope for a larger data effort. More generally, our study represents a partial equilibrium snapshot of a system that is

changing rapidly; future work might consider how other economic factors shape these differences in technology and organization.

Our next steps include merging the "births" of thousands of technological skills over the past two decades to document the causal effects of new skill launches on accumulation of complementary technologies. While the growth accounting evidence suggests an identity, the causal effect of altering the supply of complements in the market may lead to a variety of effects on AI accumulation (and accumulation of other types of capital). We have compiled the skill birth dataset and are in the process of merging it to our existing measures.

References

Abis, S. and Veldkamp, L. (2020). The changing economics of knowledge production. *Available at SSRN 3570130*.

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2020). AI and jobs: Evidence from online vacancies. Technical report, Mimeo. Massachusetts Institute of Technology.

Alekseeva, L., Azar, J., Gine, M., Samila, S., and Taska, B. (2020). The demand for AI skills in the labor market.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., and Van Reenen, J. (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, 135(2):645–709.

Babina, T., Fedyk, A., He, A. X., and Hodson, J. (2020). Artificial intelligence, firm growth, and industry concentration. *Firm Growth, and Industry Concentration (July 14, 2020)*.

Bardhan, I., Krishnan, V., and Lin, S. (2013). Research note—business value of information technology: testing the interaction effect of IT and R&D on tobin's q. *Information Systems Research*, 24(4):1147–1161.

Bartel, A., Ichniowski, C., and Shaw, K. (2007). How does information technology affect productivity? plant-level comparisons of product innovation, process improvement, and worker skills. *The quarterly journal of Economics*, 122(4):1721–1758.

Barua, A., Kriebel, C. H., and Mukhopadhyay, T. (1995). Information technologies and business value: An analytic and empirical investigation. *Information systems research*, 6(1):3–23.

Batt, R. (1999). Work organization, technology, and performance in customer service and sales. *Ilr Review*, 52(4):539–564.

Bessen, J. (2020). Information technology and industry concentration. *Journal of Law and Economics*, 63(3):531–555.

Bharadwaj, A. S., Bharadwaj, S. G., and Konsynski, B. R. (1999). Information technology effects on firm performance as measured by tobin's q. *Management science*, 45(7):1008–1024.

Black, S. E. and Lynch, L. M. (2001). How to compete: the impact of workplace practices and information technology on productivity. *Review of Economics and statistics*, 83(3):434–445.

Bloom, N., Garicano, L., Sadun, R., and Van Reenen, J. (2014). The distinct effects of information technology and communication technology on firm organization. *Management Science*, 60(12):2859–2885.

Bresnahan, T., Greenstein, S., Brownstone, D., and Flamm, K. (1996). Technical progress and co-invention in computing and in the uses of computers. *Brookings Papers on Economic Activity. Microeconomics*, 1996:1–83.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics*, 117(1).

Brynjolfsson, E. and Hitt, L. (1996). Paradox lost? firm-level evidence on the returns to information systems spending. *Management science*, 42(4):541–558.

Brynjolfsson, E. and Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic perspectives*, 14(4):23–48.

Brynjolfsson, E., Hitt, L. M., and Yang, S. (2002). Intangible assets: Computers and organizational capital. *Brookings papers on economic activity*, 2002(1):137–181.

Brynjolfsson, E. and Milgrom, P. (2013). Complementarity in organizations. *The handbook of organizational economics*, pages 11–55.

Brynjolfsson, E., Rock, D., and Syverson, C. (2021). The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–72.

Caroli, E. and Van Reenen, J. (2001). Skill-biased organizational change? evidence from a panel of British and French establishments. *Quarterly Journal of Economics*, pages 1449–1492.

Chari, V. V. and Hopenhayn, H. (1991). Vintage human capital, growth, and the diffusion of new technology. *Journal of political Economy*, 99(6):1142–1165.

Covarrubias, M., Gutiérrez, G., and Philippon, T. (2019). Explaining the rising concentration of us industries: Superstars, intangibles, globalization or barriers to entry? In *NBER Macroeconomics Annual 2019, volume 34*. University of Chicago Press.

Crouzet, N. and Eberly, J. (2018). Understanding weak capital investment: The role of market concentration and intangibles. In *prepared for the Jackson Hole Economic Policy Symposium*.

De Loecker, J., Eeckhout, J., and Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, 135(2):561–644.

Dedrick, J., Gurbaxani, V., and Kraemer, K. L. (2003). Information technology and economic performance: A critical review of the empirical evidence. *ACM Computing Surveys (CSUR)*, 35(1):1–28.

Dewan, S. and Min, C.-k. (1997). The substitution of information technology for other factors of production: A firm level analysis. *Management Science*, 43(12):1660–1675.

Eisenmann, T., Parker, G., and Van Alstyne, M. W. (2006). Strategies for two-sided markets. *Harvard business review*, 84(10):92.

Goldfarb, A., Taska, B., and Teodoridis, F. (2019). Could machine learning be a general-purpose technology? evidence from online job postings. *Evidence from Online Job Postings (October 12, 2019)*.

Hall, R. E. (2018). New evidence on the markup of prices over marginal costs and the role of mega-firms in the us economy. Technical report, National Bureau of Economic Research.

Hammer, M. and Champy, J. (2009). *Reengineering the Corporation: Manifesto for Business Revolution, A*. Zondervan.

Jin, W. and McElheran, K. (2017). Economies before scale: survival and performance of young plants in the age of cloud computing. *Rotman School of Management Working Paper*, (3112901).

Lichtenberg, F. R. (1995). The output contributions of computer equipment and personnel: A firm-level analysis. *Economics of innovation and new technology*, 3(3-4):201–218.

MacDuffie, J. P. (1995). Human resource bundles and manufacturing performance: Organizational logic and flexible production systems in the world auto industry. *ilr Review*, 48(2):197–221.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.

Milgrom, P. and Roberts, J. (1990). The economics of modern manufacturing: Technology, strategy, and organization. *The American Economic Review*, pages 511–528.

Mithas, S., Tafti, A., Bardhan, I., and Goh, J. M. (2012). Information technology and firm profitability: mechanisms and empirical evidence. *Mis Quarterly*, pages 205–224.

Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30:6338–6347.

Osterman, P. (1994). How common is workplace transformation and who adopts it? *Ilr Review*, 47(2):173–188.

Rock, D. (2019). Engineering value: The returns to technological talent and investments in artificial intelligence. *Available at SSRN 3427412*.

Saunders, A. and Brynjolfsson, E. (2016). Valuing information technology related intangible assets. *Mis Quarterly*, 40(1).

Susarla, A., Barua, A., Konana, P., and Whinston, A. B. (2011). An empirical investigation of complementarity in IT-enabled supply chain initiatives. *Available at SSRN 1975388*.

Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science*, 60(6):1452–1469.

Tambe, P., Hitt, L., Rock, D., and Brynjolfsson, E. (2020). Digital capital and superstar firms. Technical report, National Bureau of Economic Research.

Tambe, P. and Hitt, L. M. (2012). The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research*, 23(3-part-1):599–617.

Tambe, P., Hitt, L. M., and Brynjolfsson, E. (2012). The extroverted firm: How external information practices affect innovation and productivity. *Management Science*, 58(5):843–859.

Van Reenen, J. (2018). Increasing differences between firms: market power and the macro-economy.

Wu, L., Hitt, L., and Lou, B. (2020). Data analytics, innovation, and firm productivity. *Management Science*, 66(5):2017–2039.

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., Goldschlag, N., Foster, L., and Dinlersoz, E. (2021). Advanced technologies adoption and use by US firms: Evidence from the annual business survey. Technical report, National Bureau of Economic Research.

Figure 1: Sankey diagram of technology, data, network, cloud, and AI investments

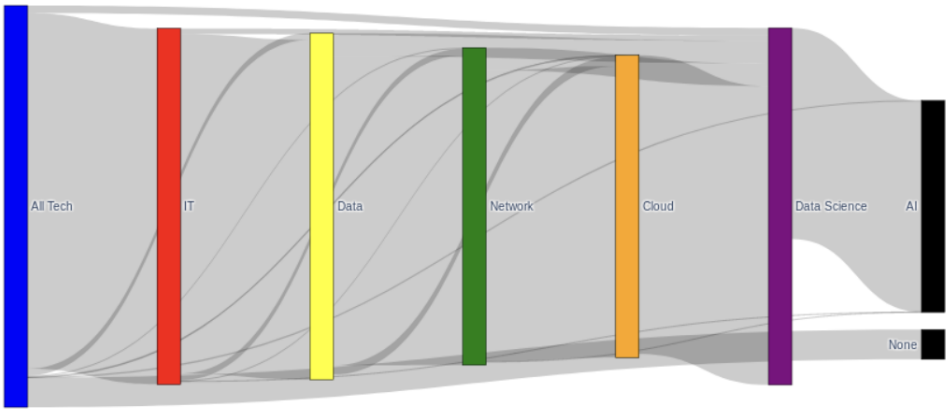


Figure Notes: This diagram depicts the flow of investments into technology and data. The blue bar is the entire Revelio sample that invests in data and technology. The red bar labeled "IT" indicates firms that invested in information technology. The yellow bar indicates firms that invest in "data", and so on, with the final group of firms investing in "AI" technologies.

Figure 2: HHI for different technology varieties over time

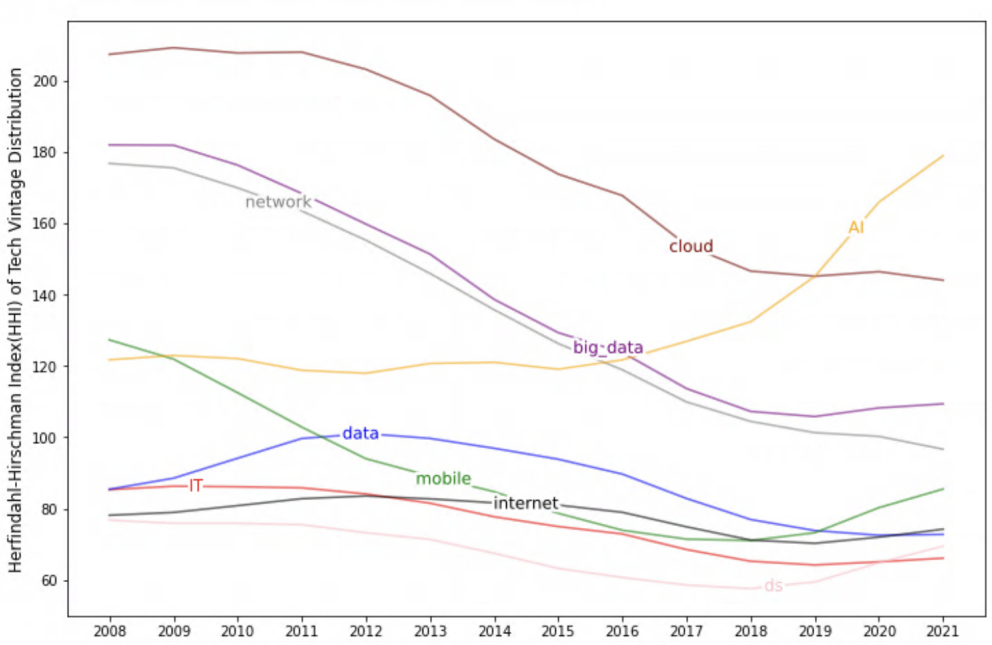


Figure Notes: In this figure, the y-axis is the HHI of skills within labeled technology varieties across firms and the x-axis is time.

Figure 3: Firms' rate of AI TFP growth against current period AI counts growth rate

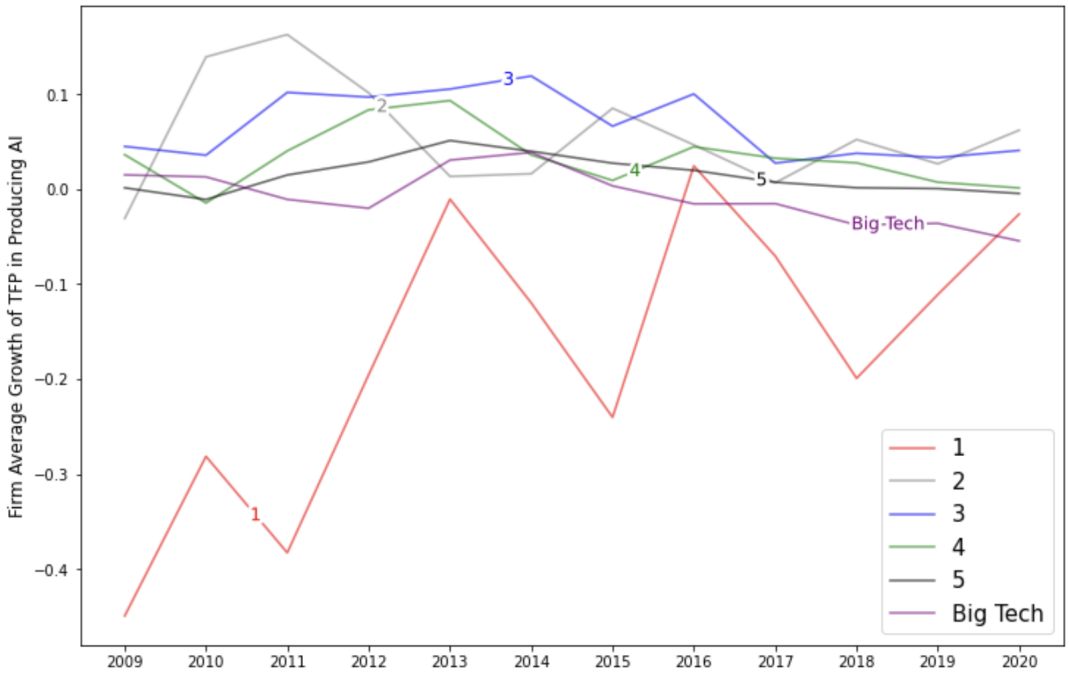


Figure Notes: In this figure, the y-axis is the rate of AI TFP growth and the x-axis is time. Firms in this figure are separated into five quintiles according to their current AI levels. 5 represents firms with the highest level of AI counts and 1 represents firms with the lowest level of AI counts. Since these are relative productivity scores, the weighted average sum of growth rates will be zero. That is, the growth of AI is the residual from a regression of AI growth on growth of other capital varieties and firm fixed effects. These values therefore strip away the grand and firm-specific static mean AI growth levels and only show within-firm change aggregated across quintiles. The largest "big tech" firms are separated out as their own group as well

Table 1: Correlation Matrix of Tech Counts

	AI	Data science	Cloud	Big_data	Internet	Mobile	Data	Network	IT
AI	1	0.745	0.436	0.645	0.706	0.766	0.498	0.292	0.628
Data science	0.745	1	0.592	0.805	0.898	0.732	0.834	0.393	0.846
Cloud	0.436	0.592	1	0.907	0.588	0.581	0.527	0.649	0.764
Big_data	0.645	0.805	0.907	1	0.822	0.771	0.752	0.689	0.926
Internet	0.706	0.898	0.588	0.822	1	0.762	0.930	0.405	0.909
Mobile	0.766	0.732	0.581	0.771	0.762	1	0.604	0.710	0.761
data	0.498	0.834	0.527	0.752	0.930	0.604	1	0.340	0.906
Network	0.292	0.393	0.649	0.689	0.405	0.710	0.340	1	0.563
IT	0.628	0.846	0.764	0.926	0.909	0.761	0.906	0.563	1

Table Notes: This table reports correlations among different firm-level technology measures.

Table 2: Balanced panel, log(MV) on log(Tech Counts) with Year Fixed Effects

	Dependent variable:								
	Log(Market Value)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log(ppegt)	0.188***	0.192***	0.188***	0.188***	0.187***	0.188***	0.189***	0.203***	0.204***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Log(oa)	0.751***	0.726***	0.728***	0.728***	0.728***	0.727***	0.728***	0.725***	0.724***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Log(AI)	0.047***	0.019***	0.031***	0.032***	0.033***	0.035***	0.035***	0.030***	0.027***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)
Log(Data science)		0.062***	0.109***	0.111***	0.116***	0.113***	0.117***	0.078***	0.080***
		(0.005)	(0.007)	(0.008)	(0.009)	(0.009)	(0.010)	(0.010)	(0.011)
Log(Cloud)			−0.070***	−0.061***	−0.059***	−0.057***	−0.056***	0.033**	0.032**
			(0.007)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
Log(Big_data)				−0.013	−0.007	−0.006	−0.005	0.028**	0.023*
				(0.012)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Log(Internet)					−0.014	−0.007	−0.0002	−0.039**	−0.038**
					(0.012)	(0.013)	(0.016)	(0.016)	(0.016)
Log(Mobile)						−0.009	−0.010	0.037***	0.035***
						(0.007)	(0.007)	(0.008)	(0.008)
Log(Data)							−0.014	0.020	0.010
							(0.015)	(0.015)	(0.017)
Log(Network)								−0.138***	−0.141***
								(0.010)	(0.011)
Log(IT)									0.023
									(0.018)
Constant	1.168***	1.127***	1.236***	1.235***	1.257***	1.243***	1.258***	1.388***	1.337***
	(0.032)	(0.032)	(0.034)	(0.034)	(0.040)	(0.042)	(0.043)	(0.044)	(0.060)
Observations	8,830	8,830	8,830	8,830	8,830	8,830	8,830	8,830	8,830
R^2	0.936	0.937	0.937	0.938	0.938	0.938	0.938	0.939	0.939
Adjusted R^2	0.935	0.937	0.937	0.937	0.937	0.937	0.937	0.939	0.939

Note: *p<0.1; **p<0.05; ***p<0.01

Table Notes: This table reports regressions of logged technology counts on logged market value. The sample is the balanced panel of firms and all models include year fixed-effects.

Table 3: Balanced panel, Log(MV) on log(Tech counts) with Year & Firm Fixed Effects

	Dependent variable:								
	Log(Market Value)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log(ppegt)	0.106***	0.100***	0.100***	0.101***	0.098***	0.097***	0.090***	0.090***	0.091***
	(0.031)	(0.031)	(0.032)	(0.032)	(0.032)	(0.032)	(0.031)	(0.031)	(0.032)
Log(oa)	0.758***	0.754***	0.754***	0.754***	0.754***	0.754***	0.752***	0.752***	0.753***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Log(AI)	0.013*	0.007	0.007	0.007	0.005	0.004	0.003	0.003	0.003
	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)	(0.008)	(0.007)	(0.007)	(0.007)
Log(Data science)		0.040**	0.040**	0.042**	0.033*	0.033*	0.018	0.018	0.019
		(0.018)	(0.018)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.020)
Log(Cloud)			−0.001	0.005	−0.009	−0.012	−0.025	−0.027	−0.021
			(0.025)	(0.028)	(0.031)	(0.031)	(0.033)	(0.035)	(0.036)
Log(Big_data)				−0.010	−0.022	−0.023	−0.029	−0.030	−0.024
				(0.021)	(0.022)	(0.022)	(0.023)	(0.024)	(0.023)
Log(Internet)					0.045	0.038	0.004	0.004	0.009
					(0.033)	(0.034)	(0.035)	(0.035)	(0.035)
Log(Mobile)						0.012	0.009	0.009	0.011
						(0.014)	(0.013)	(0.013)	(0.013)
Log(Databases)							0.097*	0.096*	0.120**
							(0.051)	(0.050)	(0.059)
Log(Network)								0.003	0.010
								(0.030)	(0.032)
Log(IT)									−0.058
									(0.070)
Constant	1.218***	1.077***	1.079***	1.083***	0.995***	1.029***	0.863***	0.861***	0.951***
	(0.164)	(0.171)	(0.177)	(0.177)	(0.178)	(0.179)	(0.192)	(0.194)	(0.206)
Observations	8,830	8,830	8,830	8,830	8,830	8,830	8,830	8,830	8,830
R^2	0.983	0.983	0.983	0.983	0.983	0.983	0.983	0.983	0.983
Adjusted R^2	0.982	0.982	0.982	0.982	0.982	0.982	0.982	0.982	0.982
Note:									*p<0.1; **p<0.05; ***p<0.01

Table 4: Within-year Regression of log(MV) on log(Tech Counts) (balanced)

	Dependent variable:												
	Log(Market Value)												
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
log(ppegt)	0.173***	0.172***	0.140***	0.158***	0.135***	0.128***	0.151***	0.129***	0.150***	0.150***	0.136***	0.146***	0.112***
	(0.023)	(0.022)	(0.022)	(0.029)	(0.020)	(0.020)	(0.019)	(0.020)	(0.019)	(0.021)	(0.022)	(0.023)	(0.024)
Log(oa)	0.781***	0.749***	0.753***	0.757***	0.780***	0.761***	0.749***	0.793***	0.757***	0.750***	0.763***	0.725***	0.765***
	(0.026)	(0.028)	(0.027)	(0.038)	(0.024)	(0.025)	(0.024)	(0.027)	(0.024)	(0.026)	(0.028)	(0.029)	(0.031)
Log(AI)	0.059***	0.041	0.038**	0.019	0.025	0.005	0.005	0.004	−0.002	0.018	0.037**	0.059***	0.068***
	(0.021)	(0.026)	(0.019)	(0.020)	(0.016)	(0.019)	(0.018)	(0.020)	(0.017)	(0.019)	(0.018)	(0.019)	(0.021)
Log(Data science)	0.064**	0.088***	0.133***	0.097***	0.086**	0.056	0.100***	0.087**	0.100***	0.136***	0.094**	0.126***	0.127**
	(0.030)	(0.025)	(0.033)	(0.036)	(0.034)	(0.042)	(0.035)	(0.038)	(0.036)	(0.044)	(0.043)	(0.045)	(0.051)
Log(Cloud)	0.038	0.071	0.030	0.089*	0.051	0.045	0.017	0.016	−0.001	−0.021	0.032	0.001	0.002
	(0.047)	(0.049)	(0.048)	(0.047)	(0.040)	(0.044)	(0.040)	(0.045)	(0.041)	(0.045)	(0.047)	(0.051)	(0.056)
Log(Big_data)	0.006	0.004	0.030	−0.018	−0.008	0.044	0.078**	0.044	0.075**	0.042	0.028	0.011	0.021
	(0.052)	(0.063)	(0.058)	(0.052)	(0.033)	(0.032)	(0.031)	(0.052)	(0.035)	(0.038)	(0.031)	(0.042)	(0.073)
Log(Internet)	−0.045	−0.022	−0.014	−0.064	−0.012	0.008	−0.013	−0.032	−0.093*	0.0003	−0.017	−0.036	−0.021
	(0.050)	(0.048)	(0.051)	(0.054)	(0.050)	(0.055)	(0.057)	(0.057)	(0.053)	(0.058)	(0.056)	(0.061)	(0.066)
Log(Mobile)	0.014	0.023	0.009	0.021	−0.008	0.025	0.030	0.045	0.053*	0.003	0.021	0.022	0.030
	(0.032)	(0.026)	(0.024)	(0.027)	(0.027)	(0.028)	(0.030)	(0.032)	(0.028)	(0.028)	(0.030)	(0.036)	(0.034)
Log(Databases)	0.063	−0.019	−0.076	0.027	0.067	0.070	0.019	0.078	0.024	−0.074	0.017	−0.058	−0.133
	(0.052)	(0.052)	(0.053)	(0.055)	(0.057)	(0.058)	(0.056)	(0.060)	(0.060)	(0.073)	(0.076)	(0.081)	(0.093)
Log(Network)	−0.109***	−0.130***	−0.081**	−0.142***	−0.114***	−0.145***	−0.137***	−0.158***	−0.159***	−0.148***	−0.158***	−0.135***	−0.155***
	(0.035)	(0.032)	(0.035)	(0.033)	(0.032)	(0.034)	(0.035)	(0.038)	(0.036)	(0.040)	(0.041)	(0.044)	(0.047)
Log(IT)	−0.054	−0.005	−0.012	0.026	−0.024	−0.041	−0.027	−0.013	0.075	0.125*	0.051	0.126	0.165*
	(0.060)	(0.062)	(0.057)	(0.063)	(0.059)	(0.058)	(0.057)	(0.066)	(0.059)	(0.069)	(0.075)	(0.081)	(0.098)
Constant	1.863***	2.323***	2.402***	1.955***	1.846***	2.239***	2.285***	2.010***	2.245***	2.216***	2.180***	2.462***	2.543***
	(0.225)	(0.320)	(0.242)	(0.245)	(0.247)	(0.255)	(0.252)	(0.277)	(0.271)	(0.272)	(0.323)	(0.322)	(0.376)
Observations	679	679	679	679	680	680	680	679	679	679	679	679	679
R^2	0.956	0.952	0.950	0.948	0.952	0.945	0.948	0.941	0.950	0.942	0.937	0.931	0.919
Adjusted R^2	0.955	0.950	0.948	0.946	0.950	0.943	0.945	0.939	0.948	0.939	0.935	0.928	0.916

Note: *p<0.1; **p<0.05; ***p<0.01

Table Notes: This table reports year-on-year regressions of logged technology counts on logged market value. The sample is the balanced panel of firms.

Table 5: Within-year regression of log(MV) on log(Tech Counts) (unbalanced panel)

	Dependent variable:												
	Log(Market Value)												
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Log(ppegt)	0.184***	0.168***	0.153***	0.161***	0.139***	0.135***	0.146***	0.124***	0.146***	0.130***	0.112***	0.130***	0.065***
	(0.021)	(0.021)	(0.019)	(0.021)	(0.017)	(0.017)	(0.017)	(0.015)	(0.014)	(0.014)	(0.016)	(0.015)	(0.017)
Log(oa)	0.797***	0.773***	0.757***	0.765***	0.794***	0.777***	0.773***	0.803***	0.760***	0.759***	0.774***	0.752***	0.820***
	(0.023)	(0.026)	(0.024)	(0.029)	(0.020)	(0.021)	(0.021)	(0.020)	(0.017)	(0.019)	(0.020)	(0.020)	(0.022)
Log(AI)	0.045**	0.022	0.006	0.019	0.032**	0.022*	0.032***	0.028***	0.020**	0.019*	0.011	0.038***	0.057***
	(0.019)	(0.021)	(0.015)	(0.015)	(0.013)	(0.013)	(0.012)	(0.010)	(0.010)	(0.011)	(0.012)	(0.012)	(0.012)
Log(Data science)	0.068**	0.095***	0.119***	0.108***	0.069***	0.074**	0.087***	0.063***	0.066***	0.087***	0.111***	0.114***	0.131***
	(0.028)	(0.023)	(0.028)	(0.030)	(0.026)	(0.030)	(0.030)	(0.020)	(0.020)	(0.022)	(0.022)	(0.021)	(0.023)
Log(Cloud)	0.048	0.094**	0.055	0.091**	0.102***	0.048*	0.064**	0.052	0.009	0.032	0.006	0.002	0.010
	(0.043)	(0.042)	(0.037)	(0.035)	(0.030)	(0.028)	(0.031)	(0.033)	(0.030)	(0.030)	(0.032)	(0.033)	(0.031)
Log(Big_data)	−0.007	0.0003	0.045	0.027	0.014	0.082***	0.069**	0.058*	0.070***	0.088***	0.094***	0.099***	0.072**
	(0.046)	(0.051)	(0.039)	(0.034)	(0.026)	(0.028)	(0.030)	(0.033)	(0.027)	(0.026)	(0.030)	(0.024)	(0.034)
Log(Internet)	−0.039	−0.053	−0.030	−0.065	−0.043	0.012	−0.023	−0.039	−0.049*	−0.106***	−0.039*	−0.059***	−0.042
	(0.043)	(0.040)	(0.042)	(0.044)	(0.039)	(0.031)	(0.030)	(0.032)	(0.029)	(0.022)	(0.022)	(0.020)	(0.027)
Log(Mobile)	0.014	0.033	0.034*	0.043**	0.009	0.018	0.029*	0.033*	0.020	0.039***	0.012	0.008	0.027
	(0.028)	(0.021)	(0.018)	(0.018)	(0.020)	(0.020)	(0.017)	(0.018)	(0.017)	(0.013)	(0.014)	(0.015)	(0.018)
Log(Databases)	0.041	−0.022	−0.069	0.006	0.056	−0.015	−0.051	0.032	0.003	0.013	0.001	−0.060	−0.127***
	(0.047)	(0.045)	(0.043)	(0.046)	(0.044)	(0.043)	(0.045)	(0.038)	(0.034)	(0.036)	(0.042)	(0.042)	(0.049)
Log(Network)	−0.102***	−0.114***	−0.079***	−0.124***	−0.089***	−0.124***	−0.120***	−0.106***	−0.088***	−0.118***	−0.101***	−0.114***	−0.128***
	(0.032)	(0.029)	(0.027)	(0.029)	(0.027)	(0.024)	(0.024)	(0.027)	(0.025)	(0.024)	(0.026)	(0.028)	(0.029)
Log(IT)	−0.043	−0.013	−0.032	−0.068	−0.118**	−0.099**	−0.067	−0.085**	−0.007	0.013	−0.012	0.054	0.075
	(0.052)	(0.053)	(0.047)	(0.048)	(0.046)	(0.045)	(0.046)	(0.043)	(0.039)	(0.043)	(0.048)	(0.049)	(0.057)
Constant	1.597***	2.138***	2.245***	2.133***	2.071***	2.482***	2.568***	2.363***	2.391***	2.497***	2.252***	2.502***	2.758***
	(0.201)	(0.287)	(0.200)	(0.195)	(0.189)	(0.180)	(0.217)	(0.180)	(0.177)	(0.181)	(0.209)	(0.232)	(0.241)
Observations	803	838	899	951	1,011	1,087	1,159	1,227	1,275	1,306	1,347	1,359	1,332
R^2	0.959	0.954	0.949	0.944	0.946	0.936	0.932	0.935	0.941	0.930	0.924	0.919	0.897
Adjusted R^2	0.958	0.952	0.947	0.942	0.945	0.934	0.930	0.933	0.940	0.929	0.922	0.917	0.895

Note: *p<0.1; **p<0.05; ***p<0.01

Table Notes: This table reports year-on-year regressions of logged technology counts on logged market value. The sample is the unbalanced panel of firms.

Table 6: Growth Accounting Productivity Regression

	Dependent variable:	
	Sales	
	balanced	unbalanced
Capital	0.276***	0.282***
	(0.037)	(0.041)
Emp	0.359***	0.351***
	(0.041)	(0.055)
Constant	0.023***	0.022***
	(0.002)	(0.002)
Observations	7,856	12,759
R^2	0.363	0.522
Adjusted R^2	0.305	0.458
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 7: AI Growth Accounting Productivity Regression (balanced panel)

	Dependent variable:							
	AI							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Capital	0.048	0.045	0.050	0.048	0.048	0.048	0.047	0.040
	(0.090)	(0.089)	(0.089)	(0.090)	(0.090)	(0.090)	(0.090)	(0.090)
Emp	0.143**	0.107	0.103	0.084	0.079	0.056	0.054	0.043
	(0.066)	(0.065)	(0.066)	(0.065)	(0.066)	(0.066)	(0.065)	(0.067)
Data science	0.302	0.265***	0.241***	0.213***	0.209***	0.171**	0.171**	0.167**
	(0.086)	(0.079)	(0.074)	(0.071)	(0.071)	(0.067)	(0.067)	(0.065)
Cloud		0.322***	0.217***	0.176**	0.162**	0.106	0.083	0.011
		(0.074)	(0.075)	(0.071)	(0.070)	(0.071)	(0.080)	(0.081)
Big_data			0.231***	0.177***	0.173***	0.149***	0.144***	0.109**
			(0.076)	(0.066)	(0.064)	(0.055)	(0.053)	(0.044)
Internet				0.239***	0.209**	0.057	0.059	−0.015
				(0.076)	(0.082)	(0.080)	(0.080)	(0.082)
Mobile					0.072	0.065	0.063	0.055
					(0.047)	(0.047)	(0.047)	(0.046)
Data						0.488***	0.475***	0.211
						(0.118)	(0.119)	(0.131)
Network							0.062	0.002
							(0.067)	(0.063)
IT								0.708***
								(0.170)
Constant	0.018***	0.033***	0.035***	0.041***	0.039***	0.049***	0.046***	0.068***
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.008)	(0.009)
Observations	7,856	7,856	7,856	7,856	7,856	7,856	7,856	7,856
R^2	0.060	0.067	0.072	0.076	0.078	0.082	0.082	0.087
Adjusted R^2	−0.026	−0.018	−0.012	−0.009	−0.007	−0.002	−0.002	0.002

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: AI Growth Accounting Productivity Regression (unbalanced panel)

	Dependent variable:							
	AI							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Capital	0.080	0.078	0.076	0.076	0.079	0.079	0.076	0.067
	(0.058)	(0.058)	(0.057)	(0.057)	(0.057)	(0.056)	(0.056)	(0.056)
Emp	0.145***	0.126**	0.114**	0.097*	0.093*	0.073	0.070	0.045
	(0.053)	(0.051)	(0.051)	(0.051)	(0.051)	(0.050)	(0.050)	(0.050)
Data science	0.357	0.336***	0.308***	0.269***	0.265***	0.230***	0.231***	0.219***
	(0.065)	(0.064)	(0.062)	(0.058)	(0.057)	(0.056)	(0.056)	(0.055)
Cloud		0.181**	0.100	0.081	0.070	0.035	0.005	−0.051
		(0.081)	(0.069)	(0.067)	(0.066)	(0.068)	(0.066)	(0.063)
Big_data			0.300***	0.254***	0.241***	0.219***	0.204***	0.135***
			(0.059)	(0.057)	(0.056)	(0.055)	(0.052)	(0.046)
Internet				0.230***	0.200***	0.144**	0.141**	0.081
				(0.067)	(0.065)	(0.066)	(0.066)	(0.064)
Mobile					0.084***	0.081***	0.079***	0.067***
					(0.023)	(0.023)	(0.023)	(0.023)
Data						0.304***	0.289***	0.097
						(0.107)	(0.105)	(0.088)
Network							0.118**	0.056
							(0.057)	(0.059)
IT								0.768***
								(0.169)
Constant	0.019***	0.028***	0.034***	0.040***	0.038***	0.045***	0.040***	0.069***
	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.008)
Observations	12,759	12,759	12,759	12,759	12,759	12,759	12,759	12,759
R^2	0.192	0.194	0.200	0.204	0.207	0.209	0.210	0.214
Adjusted R^2	0.084	0.086	0.094	0.098	0.101	0.103	0.104	0.108

Note: *p<0.1; **p<0.05; ***p<0.01

A Appendix: Mapping skills to technologies

This appendix describes how skills data were used to generate technology investment measures. In general, there are no existing taxonomies that map skill families to specific technologies (e.g. data science). We describe two approaches. First we describe an approach in which we simply manually cluster skills together into technology families. Second, we describe the use of Poincaré embeddings to uncover hierarchical taxonomies in a data-driven way.

A.1 Method 1: Manual categorization

The manually constructed mappings are shown below.

IT

"software testing","software engineering / software design","software training","software documentation","software installation / laptops","software development life cycle","embedded systems / embedded software","software","software architecture","software licensing", "software quality assurance", "software implementation","object oriented software","software deployment","open source software", "software asset management", "software project management", "software integration", "software development life cycle (sdlc)", "software development", "release management", "unix", "ftp", "object oriented design", "oop", "c++ / c", "c++ language", "microsoft visual studio c++", "visual c++","c/c++", "windows server", "windows server 2008 / windows server 2003", ".net / asp.net", "unit testing", "it governance", "sdlc", "bash", "shell", "linux", "object-oriented programming", "it audit / cisa", "assembly language", "servers", "user acceptance testing", "it", "support / server", "object-oriented programming (oop)", "continuous integration", "it infrastructure management", "operating systems","visual basic for applications (vba)", "information technology", "shell scripting / unix shell scripting", "linux system administration", "code review", "server administration", "agile testing","regular expressions", "system testing / system integration testing", "powershell", "ldap", "orm", "vb.net","linux kernel", "vdi", "ibm rational tools", "nas / enterprise storage", "smtp sap", "ivr", "ibm iseries","asp","weblogic","dos", "ibm aix","ado.net / asp.net ajax", "asp.net mvc / linq / entity framework", "vsam", "raid", "it operations"

Network

"lan-wan", "lan", "ssl", "ssl certificates", "wan", "network operations", "ip networking", "computer networking", "voice over ip (voip) / internet protocol (ip)", "network troubleshooting","network architecture", "network security","network development", "computer network operations", "wireless networking", "network administration", "san / storage area networks / netapp", "internet protocol suite (tcp/ip)", "tcp/ip", "data mapping tcp/ip protocols", "routing protocols / switching","switches / routers", "routing / qos", "wifi", "dns / dhcp", "ethernet", "wireless", "mpls", "netcool", "ccna / ccnp", "putty","wimax","snmp"

Data

"master data management", "spatial databases / web mapping", "data warehousing / etl","database administration","database", "database security", "metadata / metadata management", "oracle sql developer / oracle database", "data entry","data quality", "data acquisition", "data management", "data processing", "data integration / data warehouse architecture", "data migration", "database design","data collection", "db2", "sql", "pl/sql", "mssql / ms sql / ms sql server", "sql server management studio", "oracle sql", "sqlite", "mysql / php","performance tuning / sql tuning", "oracle pl/sql development","sql server", "microsoft sql server", "extract / transform / load (etl)","sybase", "t-sql / ssis / ssrs", "teradata", "sap hana","jsp / jdbc", "edi","sqr", "rdbms", "oracle rac", "ibm db2"

Mobile

"android","objective-c / ios development", "mobile device management", "wireless technologies", "wireless communications systems", "mobile application development", "swift / xcode", "android development / android sdk"

Mobile

"front-end","web analytics", "cascading style sheets","website development", "web services", "website updating / website administration", "web marketing", "j2ee application development / j2ee web services","web 2.0", "web applications", "responsive web design", "web development","basic html", "html + css", "html scripting", "xhtml", "html 5 / css3", "html5 / bootstrap", "html/css", "cascading style sheets (css)", "django", "css", "tomcat / jboss application server","nginx", "internet protocol suite", "node.js / react.js", "hosting", "ajax", "client/server", "search engine optimization (seo) / search engine marketing (sem)", "angularjs", "jquery / jquery ui", "fiddler", "javascript", "wireshark", "java script", "lamp","selenium / selenium webdriver","jsf / jpa", "soap","http", "backbone.js","websphere / websphere application server", ".net framework / asp.net web api", "soapui", "typescript", "jmeter"

Big Data

"distributed systems / scalability", "mongodb", "hive / apache pig", "docker / devops", "middleware", "data center", "centos / debian", "hadoop / apache spark / mapreduce","ubuntu", "server architecture", "red hat linux", "high performance computing", "vms","socket programming", "olap", "soa", "websphere mq", "multithreading", "service-oriented architecture (soa)", "ibm tivoli", "hive / apache pig"

Cloud

"microsoft azure","windows azure", "amazon services / aws", "cloud-computing", "cloud computing", "amazon web services (aws)", "cloud applications", "vmware", "openstack", "vmware esx / vmware infrastructure / vsphere"

Data Science

"data visualization", "data mining","statistical data analysis","big data", "data modeling", "data analytics / data science / big data analytics", "marketing analytics", "quantitative analytics", "analytics", "business analytics", "predictive analytics / predictive modeling", "pandas", "tableau", "nosql / redis", “numpy”, “R”, “scala”, “spark”, “julia”, “pyspark”

AI

"machine learning", "natural language processing", "image processing / computer vision", "artificial intelligence", “tensorflow”, “pytorch”, “scikit-learn"

A.2 Method 2: Using Poincaré embeddings to construct skill hierarchies

However, the scope and granularity of the skills data can benefit from methods that overlay structure onto these skills. The principal method we use to uncover structure (i.e. potential complements) from workforce skills data is the Poincaré embedding (Nickel and Kiela, 2017). Many recent advances in machine learning have been focused on representation learning; an approach that enables the mapping of complex data sets into a structured space. A key focus of representation learning is to make it easier to derive meaning from large and complex data sets.

For mapping skills to technologies, we are currently exploring using methods based in Poincaré embeddings. Poincaré embeddings differ from embeddings based in Euclidean space because they take advantage of latent hierarchical structure in the data. By embedding the data in a hyperbolic space, the use of a Poincaré embedding captures both similarity and hierarchy. Uncovering IT taxonomies from the skills data using this approach requires three steps:

1. After placing IT skills in an embedding space, a bounding box is drawn from min_x, max_x, min_y, max_y using their Poincare embeddings as shown in Figure A.4. The left panel for Figure A.4 is the embedding space for skills data posted by a large tech firm.

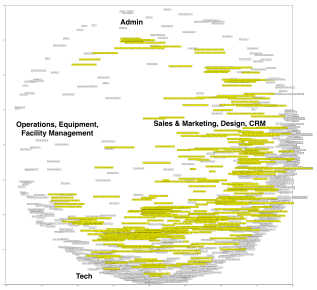
2. For all non-IT skills in the bounding box (illustrated in the right panel of Figure A.4), we compute their average distance to different IT skill groups using the distance function from

Nickel and Kiela (2017) (Equation 18).

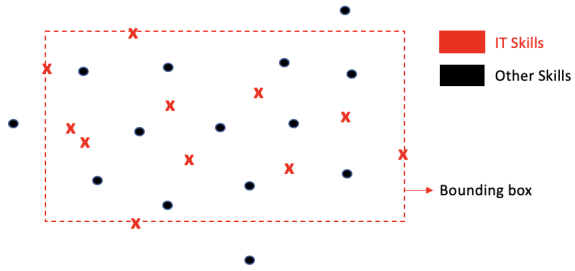
$$d(u,v) = arccosh(1 + 2\frac{\|\mathbf{u} - \mathbf{v}\|^2}{(1 - \|\mathbf{v}\|^2)(1 - \|\mathbf{u}\|^2)}) \tag{18}$$

3. The skills with the smallest average distance to all skills in the technology cluster are those we identify as complements.

Figure A.4: Using skill embeddings to identify technology complements



(a) Poincaré embeddings map for a large tech firm



(b) Bounding box to identify complements

Figure Notes: This diagram illustrates the use of a bounding box to identify which skills are complements to a pre-identified class of skills.